Propuesta de detección de datos anómalos y ruido en declaraciones juradas públicas

Rodrigo López-Pablos[1,2] y Horacio D. Kuna[3]

[1]Escuela de Posgrado y Formación Continua,
Universidad Nacional de La Matanza, Argentina
[2]Escuela de Posgrado, Facultad Regional Buenos Aires,
Universidad Tecnológica Nacional, Argentina
[3]Departamento de Informática, Facultad de Ciencias Exactas, Químicas y Naturales,
Universidad Nacional de Misiones, Posadas, Argentina
{rodrigo.lopezpablos,hkuna}@gmail.com

Resumen. Los procesos de detección de campos anómalos y con ruido son de suma utilidad para la evaluación de calidad de bases de datos de todo tipo. Estos procesos pueden tener una utilidad cívica y pública inédita si se encuentran dirigidos a la detección de valores anómalos en datos públicos. En este trabajo, se propone investigar la posibilidad de experimentación, validación y aplicación de procedimientos híbridos de detección de datos anómalos y ruido en sistemas de declaraciones juradas oficiales disponibles actualmente en Argentina.

Palabras clave: Datos anómalos y ruido, datos públicos, funcionarios públicos, declaraciones juradas, bases de datos, outliers.

1 Introducción

Los procesos de minería de datos y explotación de la información han sido escasamente usados para la resolución de problemas cívicos vinculados al sector público, los procesos de detección de datos anómalos y ruido no son una excepción en este sentido. Los datos públicos, como cualquier tipo de datos, pueden encontrarse sujetos a a nomalías y ruido como podría ser el caso de cualquier base de datos (BBDD) considerada, solo que las implicancias del descubrimiento de comportamiento corrupto en datos públicos de funcionarios públicos (FFPP) y la calidad de su confección poseen efectos profundos y de relevancia que podrían impactar en el bienestar societario dado que la corrupción condiciona el tejido social y la calidad de vida de las poblaciones. Como herramienta, la utilidad de la minería de datos para echar luz en el descubrimiento del tejido social corrupto se dirige

especialmente al ciudadano cívico y las organizaciones civiles que luchen contra la corrupción, reivindicando la utilidad de la minería de datos como herramienta en ciencias sociales, en el monitoreo y la investigación científica [1].

En la Subsección 1.1 se plantean las preguntas hipotéticas que hacen a este trabajo, la Sección 2 hace al estado de la cuestión de las técnicas de minería de datos dirigidas al comportamiento corrupto; mientras que en la Sección 3 se estudian los sistemas de DDJJ abiertas, presentándose en la Sección 4 los procedimientos de detección de anomalías propuesto, su experimentación en la Sección 5 para finalmente concluir en la Sección 6.

1.1. Preguntas de investigación

En esta propuesta de investigación se plantean las siguientes preguntas:

- ¿Es factible la experimentación, aplicación y uso de técnicas y procesos de detección de datos anómalos sobre bases de datos públicas de DDJJ oficiales como herramienta para encontrar indicios de comportamiento corrupto en la administración pública y lucha contra la corrupción?
- ¿Es factible evaluar la calidad de tales bases de datos mediante análisis de ruido y anomalías?
- ¿Qué procedimientos de detección de datos anómalos y ruido pueden ser aplicados considerando el carácter de aquellos sistemas?

2. Estado del Conocimiento

Las técnicas y proceso de minería de datos y explotación de la información han sido escasa o nulamente usados como herramienta cívica de apoyo en la lucha contra la corrupción desde la óptica pública; no obstante, encontramos literatura relevante sobre técnicas de minería dirigida a l a detección de fraude o corrupción privada, financiera y contable.

Una clasificación de su uso contra la corrupción privada, la cual se manifiesta en la forma de diversos fraudes financieros y contables, según varios autores [2,3] puede apreciarse una clasificación exhaustiva de estos últimos los cuales pueden comprender fraudes bancarios, tarjetas de crédito, blanqueo o lavado de dinero y fraude hipotecario, casos de corrupción privada compleja fraudes financieros complejos que involucran la falsificación de información corporativa, la especulación financiera, etc. A continuación una clasificación de las técnicas usadas en el campo.

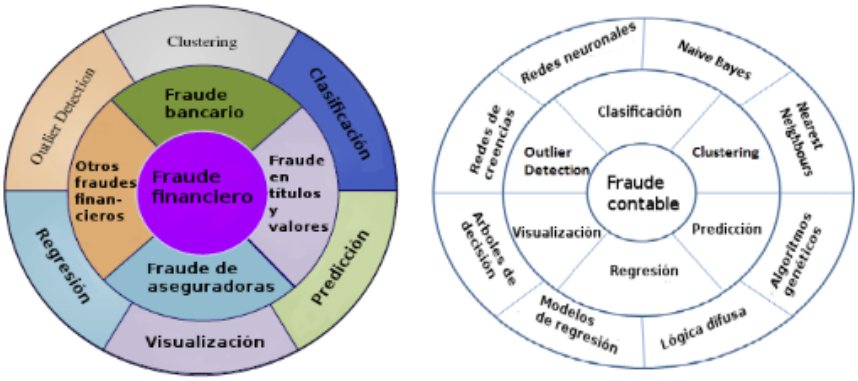

Fig. **1.** A la izquierda, la clasificación de técnicas de explotación de la información para la detección de fraude financiero [2]; y a la derecha, los algoritmos y técnicas de minería de datos para el descubrimiento de fraude contable o fiscal [3].

De la figura 1 se advierte que la utilización de estas herramientas ha sido subexplotada en el ámbito público tanto para el aprovechamiento cívico como contra la corrupción dado que no encontramos antecedentes ni literatura relevante que demuestre interés por el problema de la corrupción pública.

Así mismo, sin desmedro de las restantes técnicas de minería de datos y procesos de explotación de la información, las técnicas de detección de outliers o valores anómalos no ha sido implementada hasta ahora para la solución escenarios de fraude, y por supuesto mucho menos aún, dentro del estudio de la corrupción en el ámbito público.

3. Los sistemas de DDJJ abiertas como solución optativa en la lucha contra la corrupción pública

En Argentina, los sistemas de DDJJ son sistemas de información que se encuentran regulados por sistemas o regímenes de DDJJ; los cuales, poseen tres funciones básicas para lo cual fueron implementados [4]:

[i] Controlar la evolución patrimonial de los funcionarios de la función pública para prevenir enriquecimiento ilícito y otros delitos de corrupción.

[ii] Detectar y prevenir conflictos de intereses e incompatibilidades de la función pública.

[iii] Como mecanismo de transparencia y prevención de la corrupción pública.

En el sentido preventivo todo régimen de DDJJ de FFPP representa una herramienta que posibilita: el control del adecuado cumplimiento de las funciones públicas de los FFPP, prevenir el desvío de sus deberes éticos y corregir incumplimientos detectados.

3.1 Los sistemas y aplicativos abiertos de DDJJ

Del relevamiento exploratorio de datos públicos en Argentina, la experimentación se apoyó en el sistema interactivo de DDJJ Abiertas [5] del Diario La Nación en conjunto con las oenegés Directorio Legislativo, Poder Ciudadano y la Asociación Civil por la Libertad y la Justicia, iniciativa vigente desde 2013. De un total de 1550 DDJJ totales del sitio interactivo, 539 DDJJ son correspondientes a 99 funcionarios del poder ejecutivo, 843 DDJJ correspondientes a 313 funcionarios del poder legislativo, y 168 DDJJ correspondientes a 87 funcionarios del poder judicial; a partir de la versión actualización al 14 de abril de 2015. La estrategia de preparación de los datos para la experimentación se apoyo solamente en los bienes inmuebles de FFPP.

La BD de DDJJ abiertas presenta la siguiente estructura de atributos para cada declaración jurada de funcionario público.

```
dj.funcionario =  (ddjj_id, ano, tipo_ddjj, poder, persona_id, nombre,
                  ingreso, cargo, jurisdiccion, cant_acciones,
                  descripcion_del_bien, destino, localidad, nombre_bien_s,
                  origen, pais, porcentaje, provincia, tipo_bien_s,
                  titular_dominio, vinculo, superficiem2, val_decl,
                  valor_patrim)
```

La BD preparada para la experimentación presenta un total de 6627 tuplas con 24 atributos donde, – de 39 atributos iniciales de la BD originaria –, se descartaron 13 al encontrarse campos parcial o completamente vacíos, inconsistencias nominales en la imputación de los datos, así como atributos que pasaron a ser redundantes *a posteriori* de la homogeneización y estandarización ligados a los tres (3) atributos generados: `dj.patrimoniales (Gen)=(superficiem2, valor_patrim, val_decl)`.

Para el desarrollo experimental los valores monetarios patrimoniales en moneda extranjera fueron convertidos a pesos argentinos y actualizados por inflación `(valor_patrim)`, la superficie inmobiliaria es homogeneizada a metros cuadrados `(superficiem2)` y se categorizó el valor del inmueble declarado por el funcionario `(val_decl)` de acuerdo a su valuación fiscal relativa, pudiendo ser fiscal, subfiscal, de mercado, o no declarándose valor alguno.

4. Procedimiento híbridos de detección de datos anómalos y ruido propuestos

Los campos anómalos se definen como un dato que por ser muy diferente a los demás pertenecientes a un mismo conjunto de datos [6], *i.e.*: una base de datos contenedora de tales campos, puede considerarse que fue creada por un mecanismo diferente; lo que, en el descubrimiento de tales mecanismos, radica el conocimiento latente en cada base analizada.

Recientemente, los métodos de detección híbridos – que combinan diferentes algoritmos de distintos enfoques de aprendizaje – se han revelado como procesos así como la combinación de distintos procedimientos permite detectar outliers con un nivel de confianza mayor al 60% [7]. Los métodos híbridos de detección de anomalías, poseen la ventaja de que combinan distintas técnicas y algoritmos para un mismo propósito, por ejemplo: LOF (Local Outlier Factor) y metadatos o LOF y K-Means para BBDD numéricas así como algoritmos de inducción como C4.5, PRISM, Teoría de la Información, RB (Redes Bayesianas); y los algoritmos de clustering LOF, DBSCAN y K-Means para BBDD alfanuméricas con tipos de procedimiento tanto supervisado como no supervisado.

A continuación, (tabla 2) se describe la utilización de métodos híbridos con distintos enfoques, dependiendo del tipo de aprendizaje de los algoritmos implicados; ya sean con aprendizaje supervisado y no supervisado en la detección de anomalías y ruido. Los métodos híbridos son la mejor alternativa para lograr la mayor ganancia de información, reducción del espacio de búsqueda y optimización de los procesos [8,9,10]. Investigaciones recientes han determinado que es posible afirmar que la combinación de algoritmos de distinta naturaleza, y también la combinación de procedimientos, permite optimizar el descubrimiento de anomalías [7]; siguiendo ese paradigma, se proponen los dos siguientes procedimientos alfanuméricos para el contralor cívico.

Tabla 2. Procedimientos híbridos propuestos según entorno, algoritmos y enfoque [].

Procedimientos híbridos	Entorno	Algoritmos y técnicas	Enfoques
I	BBDD alfanuméricas con un atributo objetivo	C4.5, Teoría de la información; LOF	No supervisado; supervisado
II	BBDD alfanuméricas que no contienen un atributo objetivo	LOF; DBSCAN; C4.5; RB; PRISM; K-Means	No supervisado; supervisado

En los procedimientos híbridos I y II, donde el procedimiento I, detecta campos de outliers en bases de datos alfanuméricas conteniendo un atributo clase [11,12,7], igualmente al procedimiento II, el cual también detecta campos en bases alfanumérica solo que sin un atributo target [13,7].

Teniendo conocimiento que las DDJJ se conforman habitualmente por datos alfanuméricos, se descubre la potencialidad de aplicar en su uso los procedimientos híbridos I y II (tabla 2) correspondientes a los procedimientos híbridos de detección de datos anómalos desarrollado recientemente por [7], los cuales se identifican como idóneos por las siguientes razones:

[i] Son procedimientos desarrollados recientemente y representan el estado del arte en lo que hace a la detección de campos anómalos y ruido.

[ii] Son procedimientos óptimos para la detección de ruido en BBDD alfanuméricas como generalmente se encuentran caracterizados los datos públicos.

[iii] Son de fácil aplicabilidad y ejecución con los programas de minería de datos disponibles actualmente.

La idoneidad de estos procedimientos híbridos son propuestos como solución tentativa para un caso de contralor civil para la mejora de los datos públicos y cívicos de una población. A continuación se despliega una posible resolución y aplicación hipotética de contralor civil en base a los procedimientos alfanuméricos descriptos, sobre BBDD de DDJJ abiertas previamente tratadas.

Fig. 4. Propuesta de aplicación de procedimientos híbridos de detección de campos anómalos sobre bases de datos de declaraciones juradas alfanuméricas.[Elaboración propia]

El procedimiento propuesto (figura 4) supone aplicar los procedimientos híbridos alfanuméricos de detección de anomalías en las BBDD preparadas de DDJJ para la detección de outliers, la posibilidad de detección de falsos positivos propone la retroalimentación circular tanto en uno como otro procedimiento alfanumérico propuesto, puesto que haya o no outliers, se procede igualmente con el procedimiento restante. De esta forma, se busca evaluar la calidad de los datos públicos implicados en el análisis a primera luz, y en un análisis más profundo *a posteriori* en base a los campos y atributos detectados, los indicios de comportamiento corrupto implícito, existente como output de la información pública procesada.

5. Experimentación y discusión de la metodología propuesta

De la experimentación se procede a validación del sistema propuesto en la Figura 4 en DDJJ públicas. Puesto que no se puede afirmar qué y cuales tuplas y atributos de la BBDD de DDJJ de inmuebles son realmente válidos o no, la existencia de sospecha de anomalías se asocia a determinados parámetros económicos característicos del inmueble no triviales a su composición, así como a la calidad de los datos relevados.

5.1 Validación de los procedimientos híbridos con BBDD de DDJJ

De la ejecución experimental del procedimiento I, en su fase con aprendizaje no supervisado al valor declarado como atributo clase del algoritmo de inducción C4.5, se obtienen atributos significativos al lograr la mayor ganancia informativa (tabla 3), sobre los cuales se elaboraron 6 bins de entrada-salida simulando un sistema de información; para *a posteriori*, ejecutar los flujos de minería con algoritmo LOF, donde "∞" corresponde a la cantidad de tuplas sospechosas de contener anomalías.

Tabla 3. Bins de Entrada-Salida con outliers detectados.

Bins de Entrada – (Salida)	**Outliers detectados**	**Bins anómalos sospechosos Media o Moda(Moda)**
`superficiem2(val_decl)`	968 (∞)	38733.15(No declara)
`ano(val_decl)`	122 (∞)	2001(Mercado)
`nombre_bien_s(val_decl)`	209 (∞)	Prop. Horizontal(Fiscal)
`porcentaje(val_decl)`	252 (∞)	29.18528(Sin datos)
`val_patrim(val_decl)`	1130 (∞)	146528.3(Subfiscal)
`vinculo(val_decl)`	30 (∞)	Conviviente(Subfiscal)

Donde los bins anómalos sospechosos, se asociaron a los siguientes valores medios y modas: [i] una superficie patrimonial promedio de 38733m^2 no declarado, [ii] inmuebles de DDJJ de 2001 declaradas a valor de mercado, [iii] propiedades horizontales a su valor fiscal, [iv] tenencia de porcentaje accionario del 29,19% promedio sin valoración monetaria, [v] valor inmobiliario promedio de $146528 a valor subfiscal, e inmuebles del conviviente valuados subfiscalmente – [vi] –. Desde la teoría de la información [14], cuantitativamente, el valor patrimonial y la superficie parecen introducir mayor ruido y entropía al sistema de DDJJ.

Ejecutando el procedimiento II sin atributo clase, en la primera fase siguiendo reglas de determinación de outlier [7] se aplican LOF-DBSCAN y unión de algoritmos de clasificación C4.5-RB-PRISM donde se detectaron 2531 tuplas

sospechosas de contener anomalías, *i.e.*, un 38,19% de un total de 6627 tuplas, se conforma una BBDD de outliers para la siguiente fase la cual se transforma *a posteriori* para aplicar K-Means en dos agrupamientos, observándose los siguientes resultados.

Tabla 4. Distancia del centroide para cada atributo.

Atributo(id) (Cluster_0)	**Valor distancia (Cluster_0)**	**Valor promedio** (Cluster_1)
ddjj_id(1)	**1.841**	1.079
ano(2)	**1.518**	1.079
superficiem2(22)	**2.033**	1.079

Al ejecutar el algoritmo K-Means, el centroide más alejado contuvo los atributo superficie patrimonial, año e identificación de DJ; donde el primero – `superficiem2` – resulta el atributo más alejado así como el más sospechoso de contener campos anómalos.

5.2 Discusión de los resultados de la experimentación

El ruido anómalo producido por los atributos superficie, valuación, nombre del bien, porcentaje accionario, año y vínculo (tabla 3) vislumbra un sistema de información que expone las anomalías de los FFPP a la hora de valuar y declarar sus inmuebles frente a la ciudadanía; por otra parte, el atributo superficie presenta una fuente importante de dispersión que podría conllevar indicios de inmuebles extremadamente grandes junto a inmuebles irrisoriamente pequeños para ser considerados como tal.

Del primer análisis de inducción con atributo clase – procedimiento I –, mediante algoritmo C4.5, también se observaron las siguientes reglas:

[i] Los inmuebles menores a 6500m^2 tienden a ser declarados a valor subfiscal.

[ii] los inmuebles mayores a 6500m^2 – entre mediados de 2005 y 2012 – tienden a no declararse, ocurriendo lo mismo con cocheras, campos, terrenos, parcelas, lotes y propiedades horizontales sin especificación mayores a 6500m2 y anteriores a 2005.

[iii] cuando el funcionario posee una casa, y una participación accionaria superior al 46,3% [100%; 37,9%) o inferior al 37,9 (37,9% ; 0%] prefiere no declarar su valuación, pero tiende a declararla subfiscalmente cuando posee un rango accionario entre el 46% y el 38% (46,3%; 37,9%).

[iv] Si el FP posee un departamento, entre mediados de 2003 y mediados de 2005, con superficie menor o igual a 26m^2, tiende a ser declarado a su valuación fiscal, mientras no se declara cuando es anterior a 2003. No obstante, los departamentos mayores a 26m^2 simplemente tienden a no declararse.

[v] Los locales comerciales de los FFPP que poseen una valuación mayor a $76493 [∞, $76493) o i nferiores a $10615 ($10615, 0] tienden a ser declarados a su valor de mercado; a excepción de los locales con rango patrimonial entre [$76493, $10615], puesto que en este caso también se tenderá a ser declarado a su valor de mercado, si y solo si, el inmueble se encuentra inscripto a nombre de su cónyuge, y no precisamente a nombre del oficial político, caso en el cual se tenderá a no declararlo.

6. Conclusiones y trabajos futuros

En este trabajo se constituyeron y combinaron procesos híbridos detectores de outliers y ruido configurados para trabajar con BBDD alfanuméricas en BBDD de DDJJ de bienes inmuebles, como caso inédito de uso de técnicas y procesos de minería de datos como herramienta contra la corrupción pública, poniéndose de relieve la potencialidad de los procesos de detección de anomalías al momento de evaluar la calidad de las BBDD públicas por un lado, y el descubrimiento de información sospechosamente portadora de comportamiento corrupto por el otro.

Los atributos superficie patrimonial, valor patrimonial, porcentaje accionario, nombre del bien y año son los que aportaron más entropía al sistema de DDJJ comprometiendo la calidad de la BBDD; dada su baja densidad relativa, la superficie del inmueble es el atributo más sospechoso de contener datos anómalos. Así mismo, las posibles anomalías detectadas; además de revelar inconsistencias en la BBDD, podrían develar indicios de comportamiento corrupto respecto la valuación fiscal inmobiliaria declarada por FFPP, puesto que las características del inmueble podrían condicionar la valuación patrimonial del mismo a su valor fiscal, subfiscal, o simplemente evitando declarar su valor. En este sentido, tanto las reglas de comportamiento como los bins entrada-salida podrían servir de estrategia de investigación cívica y contable en el combate a la corrupción pública, la evasión y la elusión impositiva en FFPP; lacerante, al sopesarse la ejemplaridad ideal del funcionario respecto la comunidad representada.

Futuros trabajos de investigación podrían contemplar la elaboración de variaciones algorítmicas en los procedimientos alfanuméricos propuestos; no descartándose la aplicación de variantes a l os procesos alfanuméricos propuestos así como el de procesos de explotación de la información ajenos a la detección de datos anómalos y ruido sobre las mismas BBDD de DDJJ.

7. Referencias

[1].Ransom J.: Replicating Data Mining Techniques for Development: A Case of Study of Corruption, Lund University, Master Thesis, Master of Science in International Development and Management, http://lup.lub.lu.se/record/3798253/file/3910587.pdf, (2013).

[2].Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X.: The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559-569 (2011).

[3].Sowjanya, S., J yotsna G.: Application of Data Mining Techniques for Financial Accounting Frad Detection Scheme. International Journal of Advanced Research in Computer Science and Software Engineering, Noviembre, 3(11), 717-724 (2013).

[4].Gómez N., Bello M. A.: Ética, transparencia y lucha contra la corrupcion en la administracion publica, Manual para el ejercicio de la funcion publica, 1ra ed.: Oficina Anticorrupción, Ministerio de Justicia y Derechos Humanos de la Nación, Mayo, CABA, http://www.anticorrupcion.gov.ar/documentos/Libro%20SICEP%202da%20parte.pdf, (2009).

[5].DD.JJ. Abiertas: LNData. Actualizado al 13/1/2014, CABA, http://interactivos.lanacion.com.ar/declaraciones-juradas/, (2015).

[6].Hawkins, D. M.: Identification of outliers, Chapman and Hall., 11, London (1980).

[7].Kuna H.: Procedimientos de explotación de la información para la identificación de datos faltantes con ruido e incosistentes, Tesis doctoral, Universidad de Málaga, Marzo (2014).

[8].Kuna, H., García Martínez, R., Villatoro, F.: Identificación de Causales de Abandono de Estudios Universitarios. Uso de Procesos de Explotación de Información. Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología, 5, 39--44 (2009).

[9].Kuna, H., García-Martínez, R. Villatoro, F.: Pattern Discovery in University Students Desertion Based on D ata Mining. In Advances and Applications in Statistical Sciences Journal, 2(2): 275–286 (2010).

[10].Kuna, H., Pautsch, G., Rey, M., Cuba, C., Rambo, A., Caballero, S., Steinhilber, A., García-Martínez, R., Villatoro, F.: Avances en procedimientos de la explotación de información con algoritmos basados en la densidad para la identificación de outliers en bases de datos. Proceedings XIII Workshop de Investigadores en Ciencias de la Computación. Artículo 3745 (2011).

[11].Kuna, H. , Pautsch, G., Rey, M., Cuba, C., Rambo, A., Caballero, S., García-Martínez, R., Villatoro, F.: Comparación de la efectividad de procedimientos de la explotación de información para la identificación de outliers en bases de datos. Proceedings del XIV Workshop de Investigadores en Ciencias de la Computación, 296--300 (2012b).

[12].Kuna, H., Pautsch, G., Rambo, A., Rey, M., Cortes, J., Rolón, S.: Procedimiento de explotación de información para la identificación de campos anómalos en base de datos alfanuméricas. Revista Latinoamericana de Ingeniería de Software, 1(3): 102--106 (2013b).

[13].Kuna, H., García-Martínez, R., Villatoro, F.: Outlier detection in audit logs for application systems. Information Systems (2014).

[14].Ferreyra M.: Powerhouse: Data Mining usando Teoría de la información, (2007).